PATRIC update

BRC3

February 2006

PATRIC's Pathogens

- Bacteria:
  - *Brucella*
  - *Rickettsia*
  - *Coxiella burnetii*
- Viruses:
  - Coronaviruses (SARS)
  - Caliciviruses
  - Hepatitis A viruses
  - Hepatitis E viruses
  - **Rabies viruses ⇒ Lyssaviruses** (expanded scope)

Curation Concepts

- **Nucleotide-level Curation**
  - CDSs and RNA gene calls
  - Ribosome binding sites and start site correction
  - Pseudogenes and other physical features
- **Protein-level Curation**
  - Functional assignment, classification
  - Structural properties and features
- **Automated and Manual Curation**
- **Reference and Associated Genomes**

Curation Progress

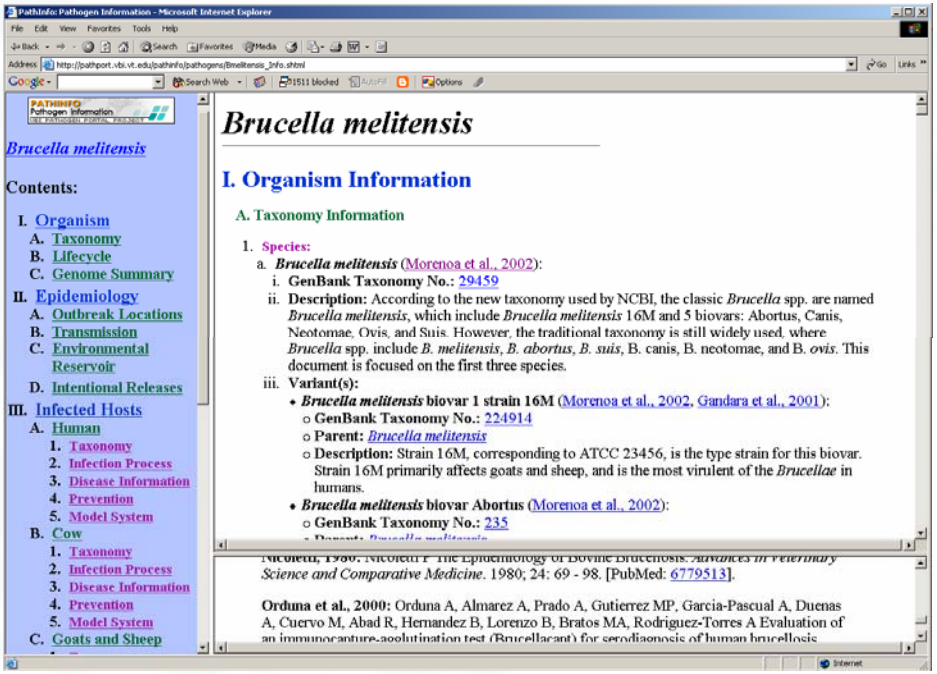
- All bacterial and viral **reference genomes** have received **nucleotide-level manual** curation
  - latest release: **Dec 22, 2005**
- **Protein-level automated** curation has been done on all bacterial reference genomes (not public yet)

Summary of Nucleotide-level Curation for Reference Genomes

Pathosystem	#genes GenBank entry	Start-site changes	Frameshifts	Premature stops	Genes deleted	Genes added
Brucella	3198	961	100	50	6	350
Coxiella	2052	74	1	0	0	124
Rickettsia	835	32	22	1	0	74

	No. of RGs	Total genes/genome	1st pass check	New annotations
Calicivirus	13	2,3	Y	
Coronavirus	16	14	Y	SARS (SZ-3)
Hepatitis A	1	2	Y	
Hepatitis E	5	3,4	Y	
Lyssavirus	1	5	Y	

Literature Curation: PathInfo



Curated Information

- Organism background
- Epidemiology
- Pathogenesis
- Experiments

Pathosystem	
Bacteria	**Status**
Brucella	Available
Coxiella	Available
Rickettsia	Available
Viruses	
Calicivirus	Ready, awaiting release
Coronavirus	in queue
Hepatitis A	in queue
Hepatitis E	Ready, awaiting release
Lyssavirus	In progress

PathInfo Status for the Pathogens of Other BRCs

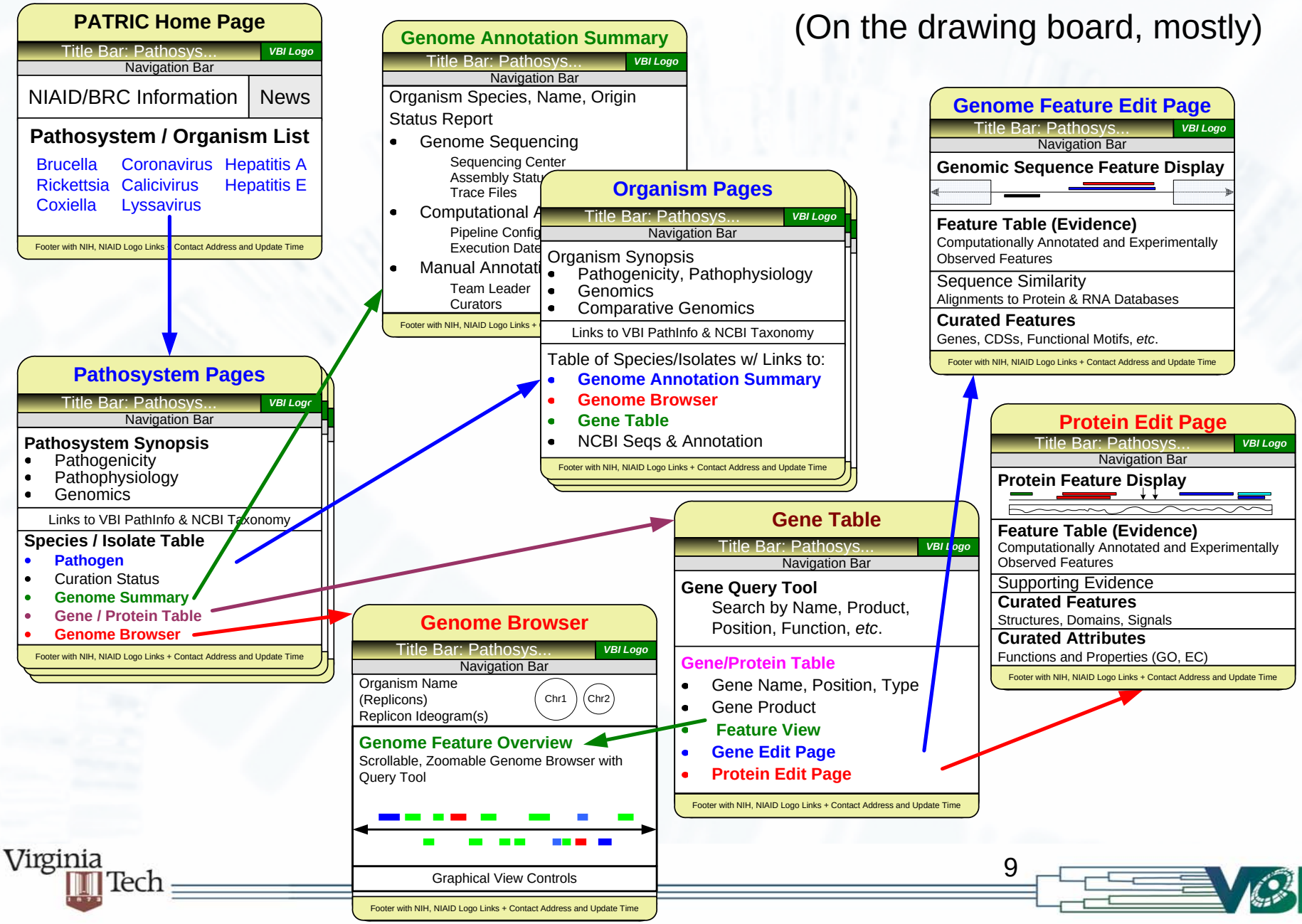
Eukaryotic pathogens		
BRC	**complete**	**to be completed**
Apicomplexan	2	1
BioHealthBase	0	1
Pathema	0	1
Viruses		
	complete	**to be completed**
VBRC	16	4
BioHealthBase	0	1
Bacteria		
	complete	**to be completed**
Pathema	3	2
NMPDR	0	4
ERIC	4	1
BioHealthBase	1	0

Progress: Software Development

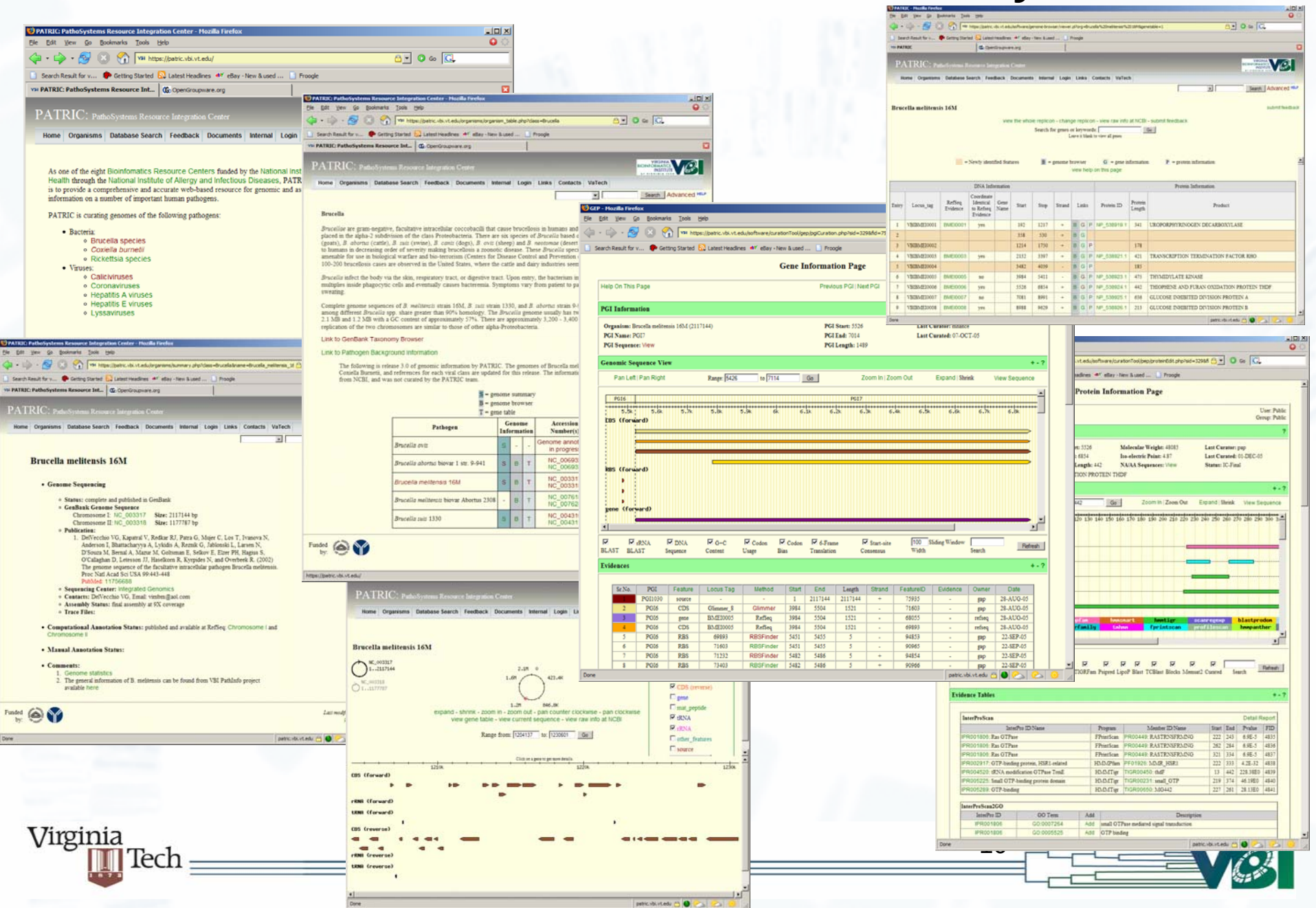
- **Curation Infrastructure**
  - Nucleotide-level curation: pipeline + edit page
  - Protein-level curation: pipeline + edit page
- **Improvements to External Website**
  - User-friendly database query interface
  - Enhanced gene/protein table display highlighting corrections to published annotations

Curation Infrastructure: February '05

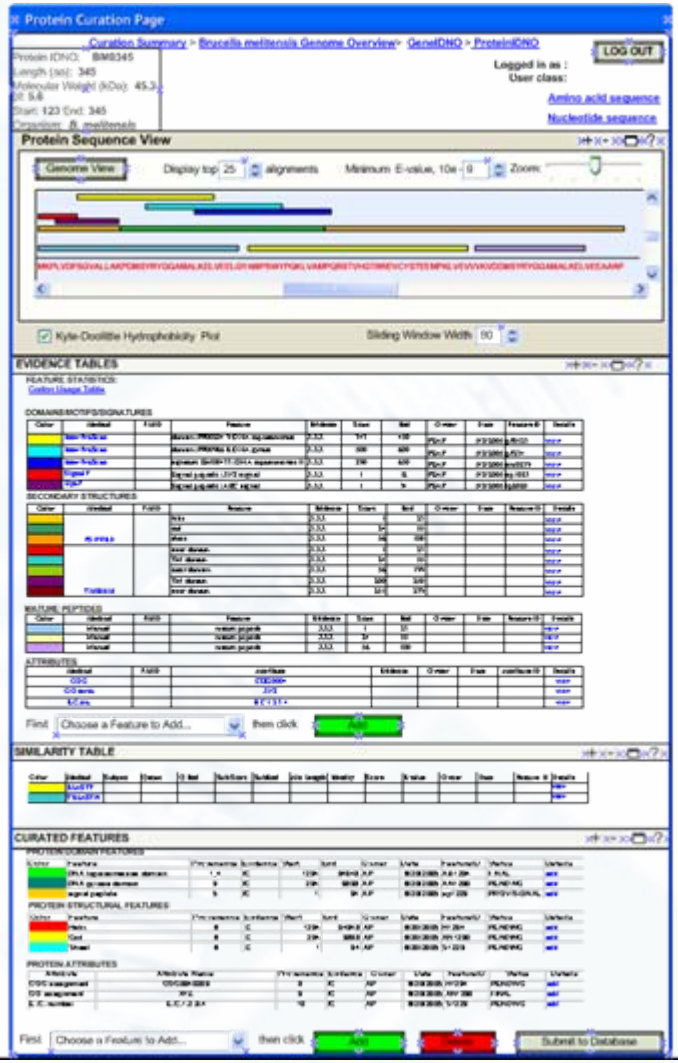
(On the drawing board, mostly)



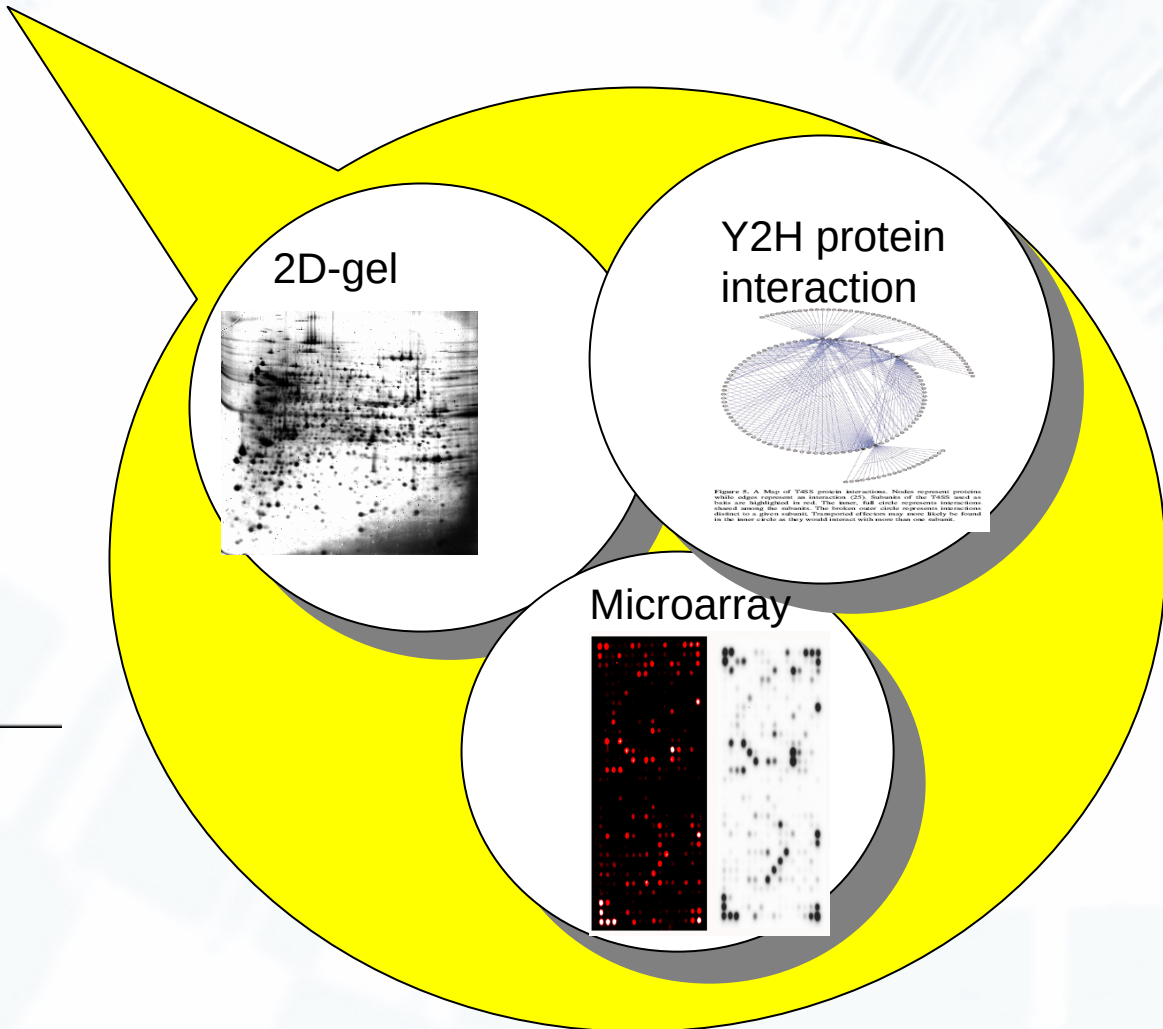
Curation Infrastructure: February ’06



Virginia Tech



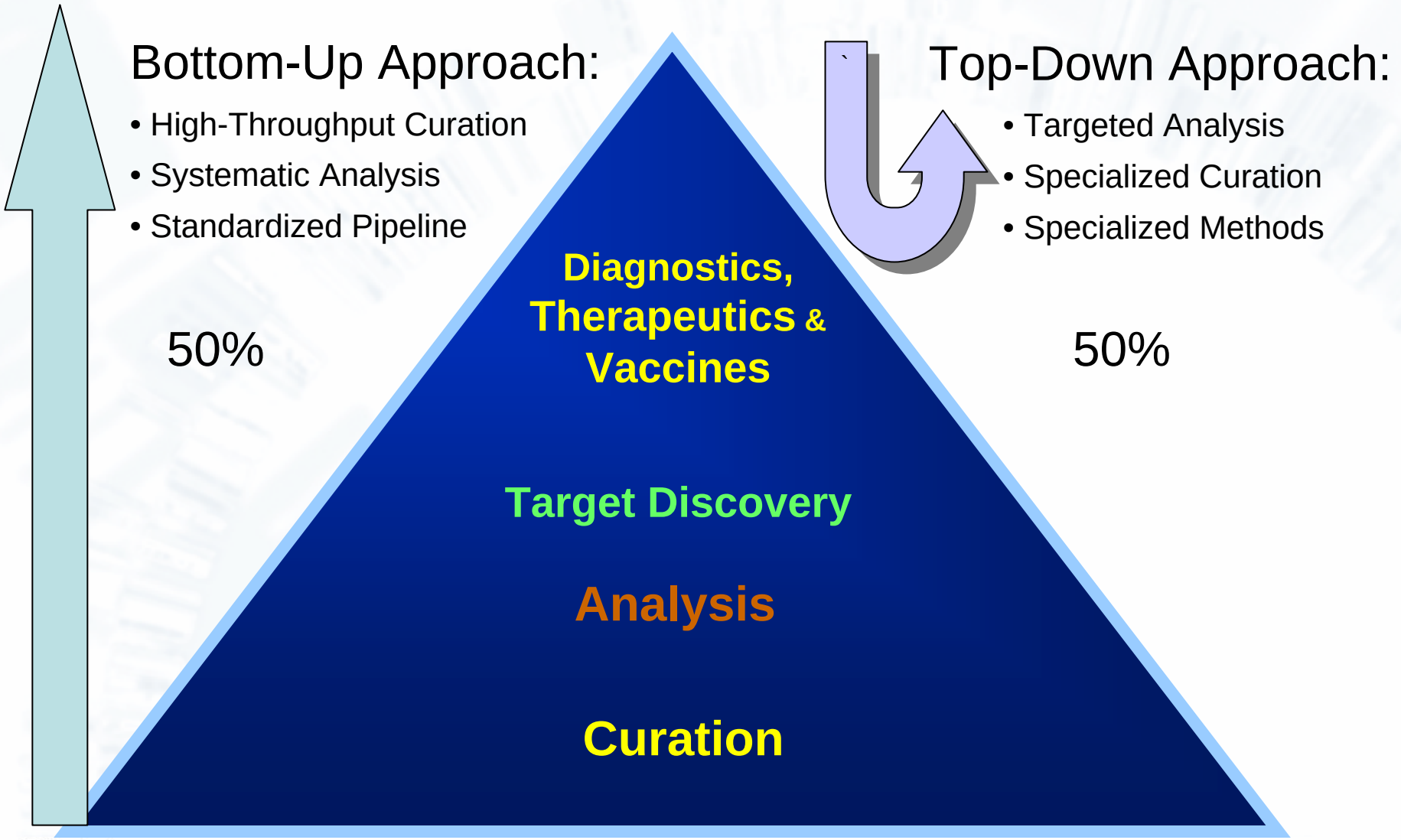
In the works: links from protein edit page to proteomics data



Progress: Databases

- Updated PATRIC database schema to GUS 3.5 for all database instances
- Various improvements and additions to database infrastructure

Curation: Two Parallel Approaches



Top-Down Approach: Targeted Discovery of Countermeasure Candidates

- Community involvement
  - Goals defined by end users (organism experts and community they represent)
  - Experts' involvement at all stages
  - Commitment from end users for validations/follow-up
- Problem-driven curation and analysis

Top down approach implementation: “Special Projects”

- *Brucella*
- *Rickettsia*
- Lyssavirus

Brucella Special Project

- Project Goals
  - Identify virulence factors by comparative genomics
  - Identify functional polymorphisms
  - Obtain diagnostic markers
  - Incorporate methodology into genome annotation pipeline
- Collaborators
  - Stephen Boyle (PATRIC Organism Expert, VirginiaTech)
  - Yongqun “Oliver” He (University of Michigan)

some preliminary results

Newly-Identified Genes
in published *Brucella* genomes

"Newly-identified genes":

- Genomic regions not previously associated with coding sequences
- …which have full-length alignments with known or predicted proteins in closely related genomes

	Newly-identified Genes	With assigned function
B. suis 1330	50	14
B. abortus 2308	129	14

Virulent vs. Attenuated Strain Comparison

- VBI has access to two unpublished *Brucella abortus* strains: 9-941 and S19
  - 9-941: parent strain
  - S19: attenuated mutant of 9-941 that developed spontaneously and is used as vaccine strain

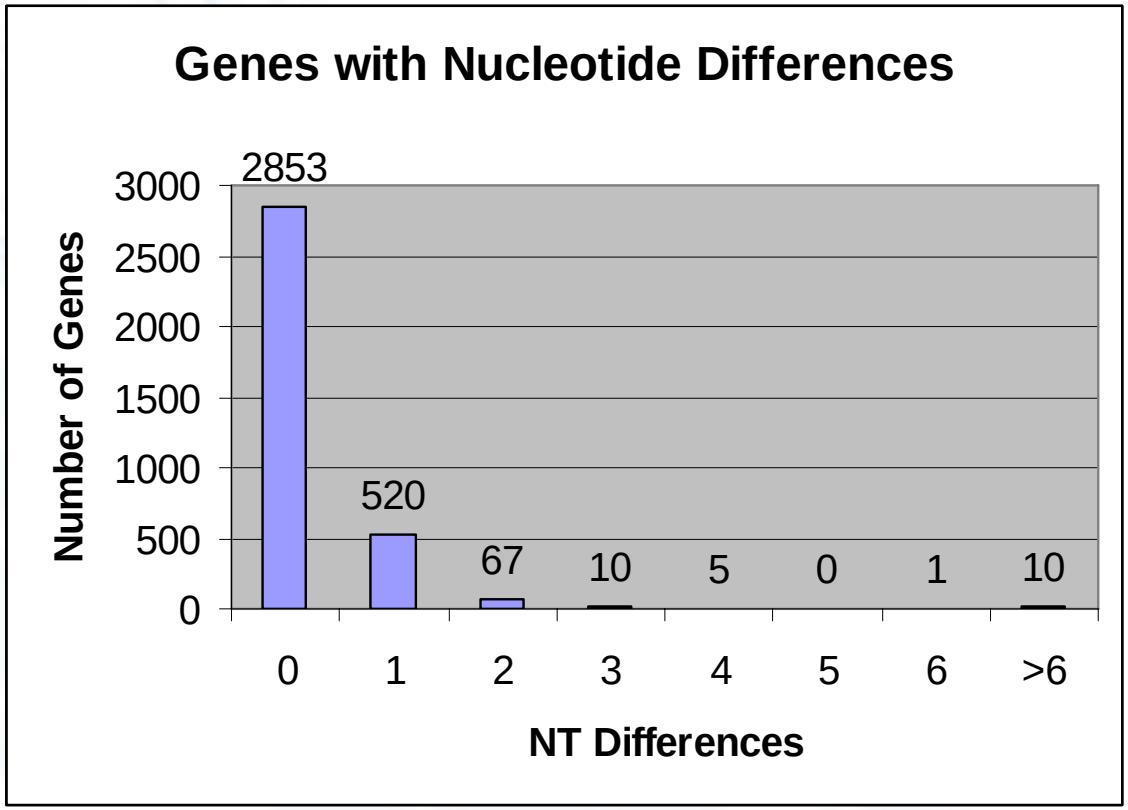
Frameshifts and Premature Stops in *Brucella abortus* S19 ORFs

S19 Analysis	Size	Premature Stop	Frameshift	Both
Chr I	2.12 Mb	34	477	10
Chr II	1.16 Mb	31	320	7

In progress: filter to distinguish

1) sequencing errors
2) mutations shared with 9-941
3) mutations specific to S19

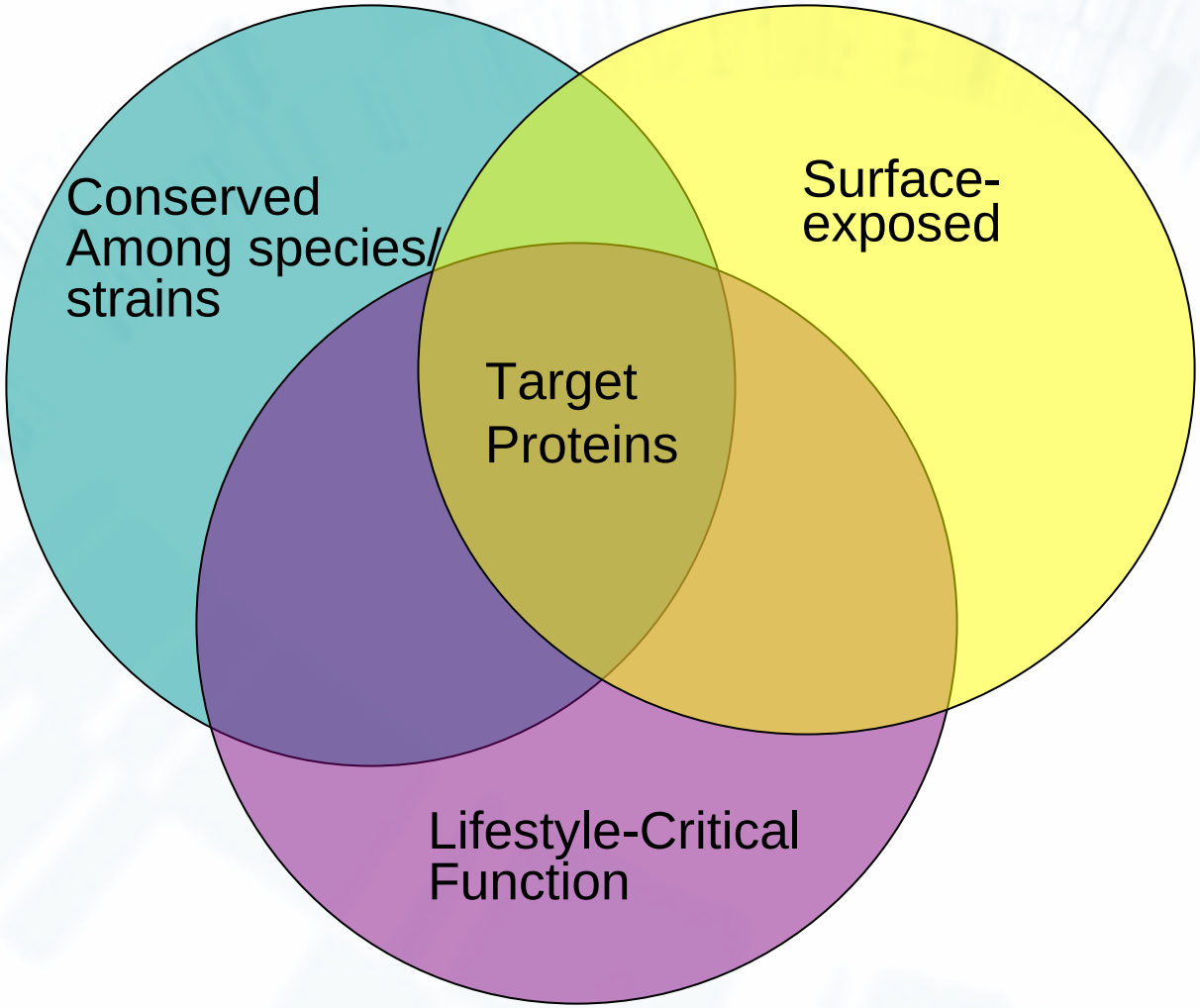
Polymorphisms (Substitutions and Indels) between Orthologs in S19 and 9-941

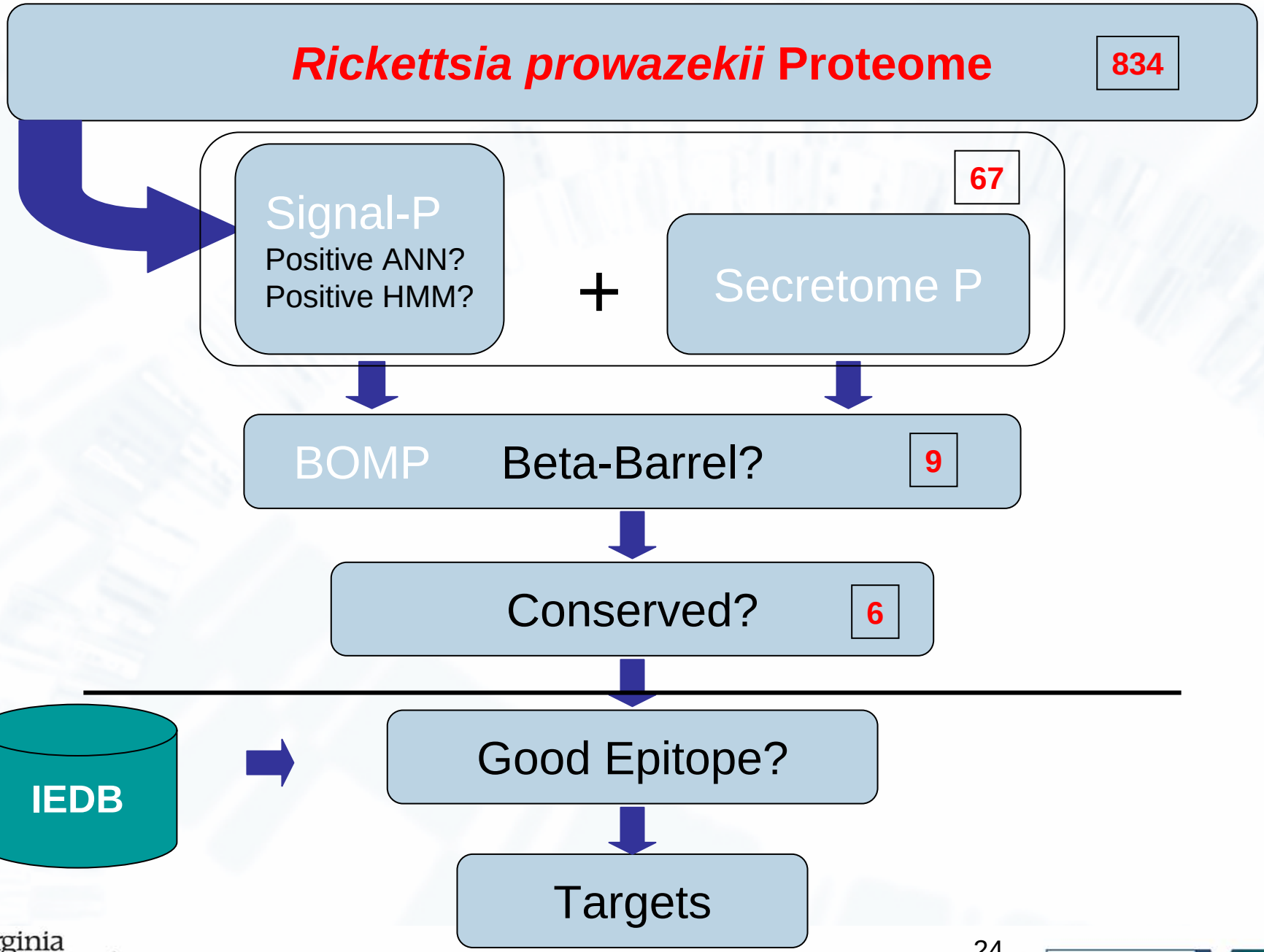


Rickettsia Special Project

- Collaborator: Abdu Azad (PATRIC Org. Expert, U. of Maryland)

Finding Good Vaccine Targets




Rickettsia prowazekii Proteome
834
Signal-P
Positive ANN?
Positive HMM?
+
Secretome P
67
BOMP
Beta-Barrel?
9
Conserved?
6
IEDB
Good Epitope?
Targets

Vaccine Target Candidate Proteins

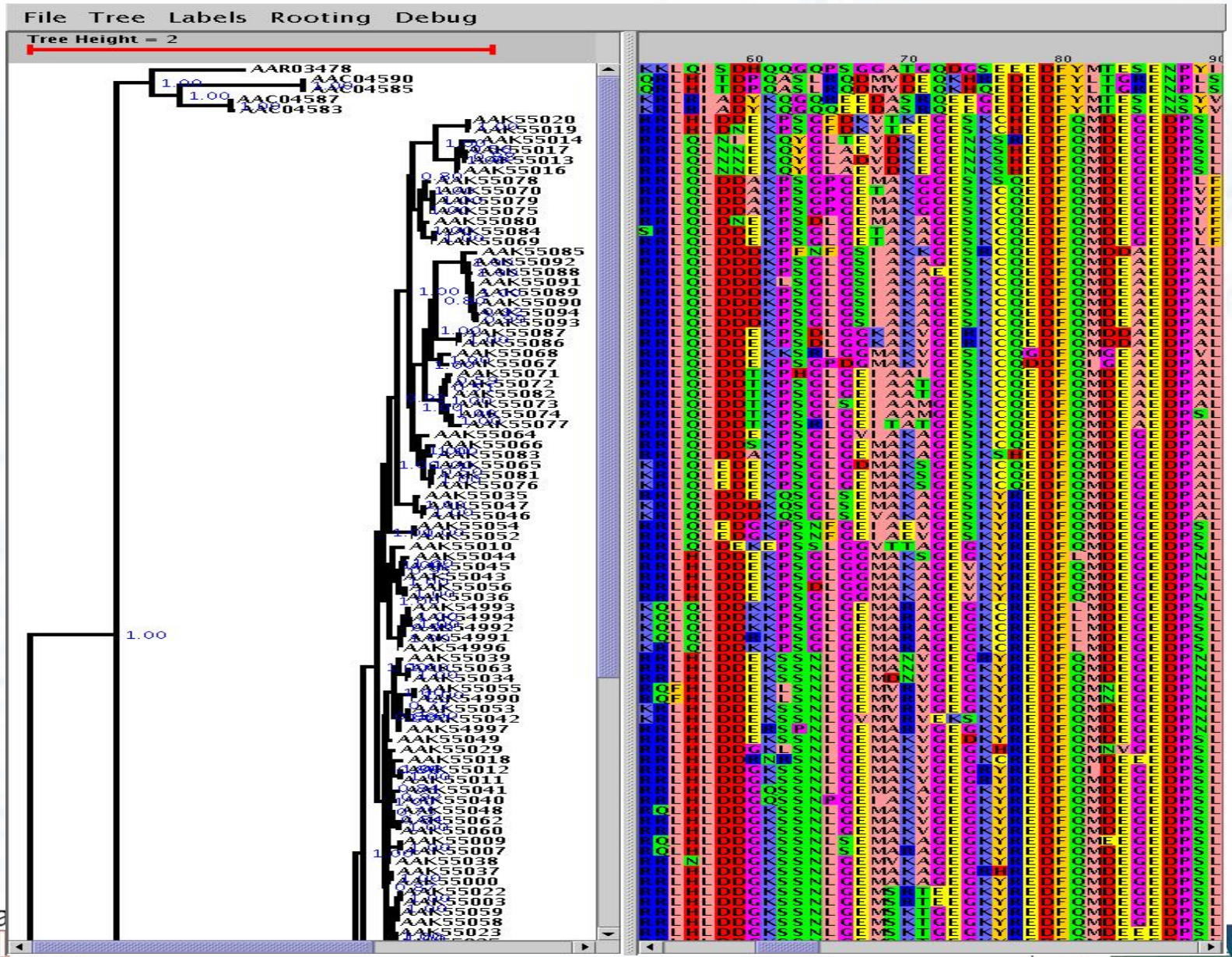
- The six genes are annotated as **hypothetical proteins**
- Highly conserved among 7 annotated Rickettsial genomes
- Further analysis: the coded proteins seem to share common features/ motifs characteristic of **autotransporters, competence factors and secretion apparatus** proteins

Lyssavirus Special Project

- Goal: develop an automated **phylogenetic classification system** (phylotyping) for Lyssaviruses
- Collaborators
  - Charles Rupprecht (PATRIC OE, CDC)
  - Alan Dickerman (VBI)

Phylotyping

- Tool that will allow submission of sequences for multiple sequence alignment and generation of phylogenetic tree (web-based)
- Methods: MUSCLE + Mr.Bayes + VBI programs
- Tree Viewer juxtaposes tree view with MSA


File Tree Labels Rooting Debug
Tree Height = 2
AAR03478
AAC04590
AAC04585
AAC04587
AAC04583
AAK55020
AAK55019
AAK55014
AAK55017
AAK55013
AAK55016
AAK55078
AAK55070
AAK55079
AAK55075
AAK55080
AAK55084
AAK55069
AAK55085
AAK55092
AAK55088
AAK55091
AAK55089
AAK55090
AAK55094
AAK55093
AAK55087
AAK55086
AAK55068
AAK55067
AAK55071
AAK55072
AAK55082
AAK55073
AAK55074
AAK55077
AAK55064
AAK55066
AAK55083
AAK55065
AAK55081
AAK55076
AAK55035
AAK55047
AAK55046
AAK55054
AAK55052
AAK55010
AAK55044
AAK55045
AAK55043
AAK55056
AAK55036
AAK54993
AAK54994
AAK54992
AAK54991
AAK54996
AAK55039
AAK55063
AAK55034
AAK55055
AAK54990
AAK55053
AAK55042
AAK54997
AAK55049
AAK55029
AAK55018
AAK55012
AAK55011
AAK55041
AAK55040
AAK55048
AAK55062
AAK55060
AAK55009
AAK55007
AAK55038
AAK55037
AAK55000
AAK55022
AAK55003
AAK55059
AAK55058
AAK55023
1.00
60
70
80

General plans for near future

- Steady state curation
  - Next release in June 2006
- Various improvements to infrastructure (including “automated curation”)
- LANL sequence annotation
- Special projects

Organization of Relevant Meetings by VBI

- International Symposium on the Comparative Biology of Alpha-Proteobacteria
  - **April 26-29**, Blacksburg
  - (*Brucella* and *Rickettsia* are alphas)
- Computational Genomics ’06
  - **October 28-31** (tentative), Baltimore
  - 3 main themes
    - **Infectious diseases**, automated annotation, and biological networks

PATRIC

PIs
B. Sobral
J. Setubal

Executive Committee
O. Crasta
M. Czar (project management)
R. Kenyon (project management)
A. Purkayastha (curation)
E. Snyder (bioinformatics)
R. Will (software)

Curators
C. Dharmanolla (literature)
V. Dongre (Hep E)
M. Hance (Rickettsia and Lyssavirus)
D. Jukneliene (Coronavirus)
L. Mackasmiel (Calicivirus)
J. Shallom (Coxiella and Hep A)
G. Yu (Brucella)

Software Developers
N. Kampanya (web design and visualizations)
J. Lu (database architect and administrator)
M. Shukla (genome curation interfaces)
J. Soneja (annotation pipelines)
F. Zhang (lit. curation, web navigation and queries)

Organism Experts
A. Azad (U. Maryland, Coxiella and Rickettsia)
S. Baker (Loyola U., Coronavirus)
S. Boyle (VT, Brucella)
Y. He (U. Michigan, Brucella)
Y. Khudyakov (CDC, Hep A)
XJ Meng (VT, Hep E)
C. Rupprecht (CDC, Lyssavirus)
J. Vinje (CDC, Calicivirus)

Collaborators
J. Gabbard and D. Hix (VT, usability engineering)
N. Ramakrishnan (VT, data mining)

Thank You

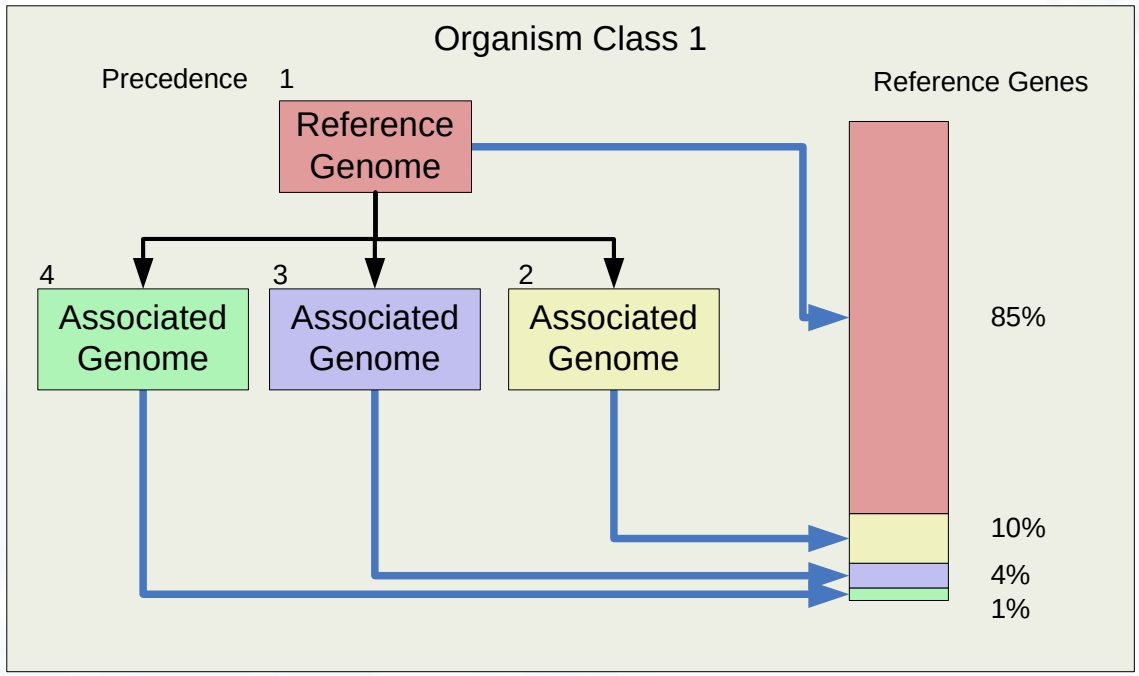
What is a Reference Genome?

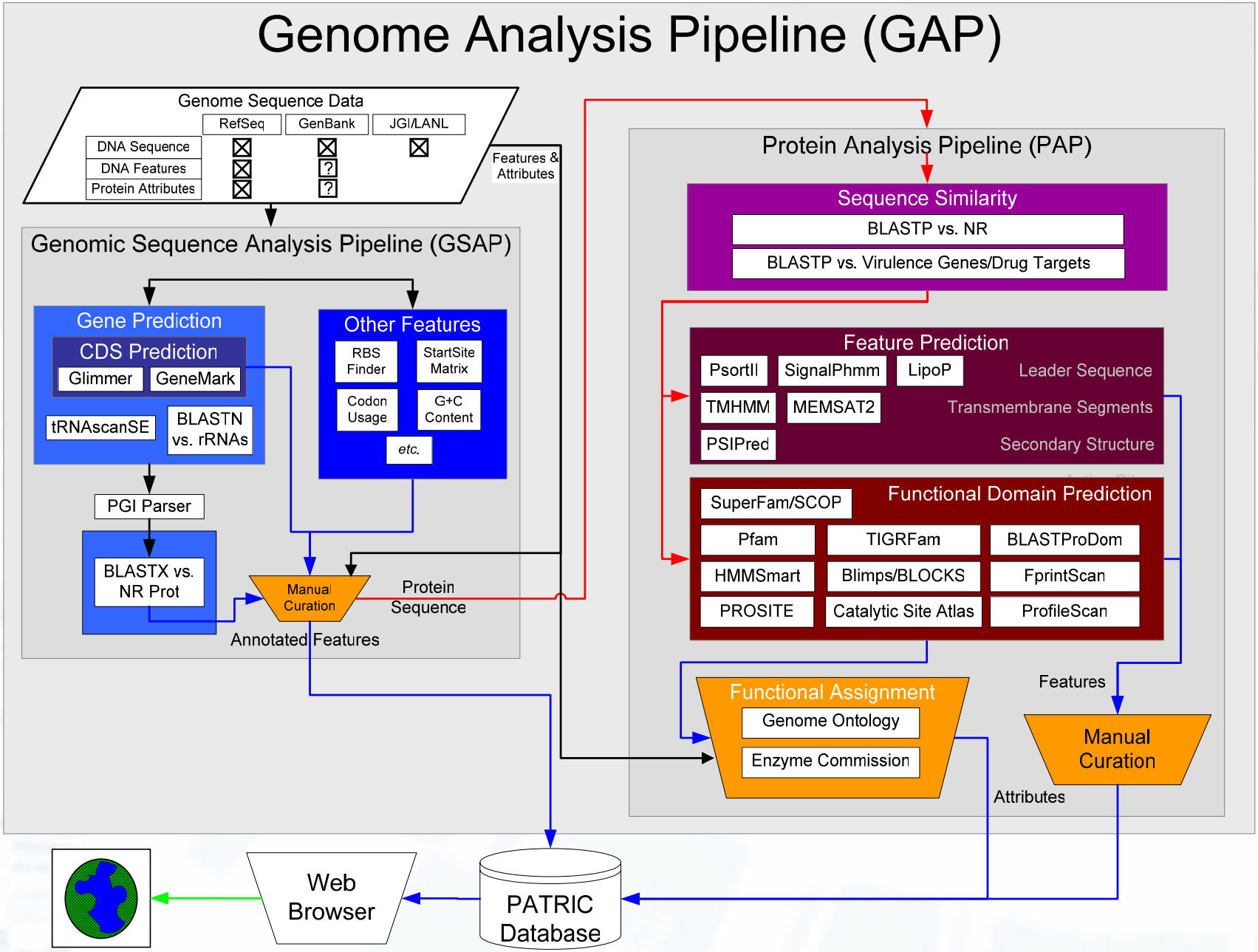
-Genome sequence of the type species of a genus.

-A lab strain that has been extensively characterized.

-A representative of a phylogenetic subgroup (community choice)

The Reference Genome annotation is the starting point to create a Reference Gene Set for that Pathosystem.




Genome Analysis Pipeline (GAP)
Genome Sequence Data
RefSeq
GenBank
JGI/LANL
DNA Sequence
DNA Features
Protein Attributes
Features & Attributes
Genomic Sequence Analysis Pipeline (GSAP)
Gene Prediction
CDS Prediction
Glimmer
GeneMark
tRNAscanSE
BLASTN vs. rRNAs
PGI Parser
BLASTX vs. NR Prot
Other Features
RBS Finder
StartSite Matrix
Codon Usage
G+C Content
etc.
Manual Curation
Protein Sequence
Annotated Features
Protein Analysis Pipeline (PAP)
Sequence Similarity
BLASTP vs. NR
BLASTP vs. Virulence Genes/Drug Targets
Feature Prediction
PsortII
SignalPhmm
LipoP
Leader Sequence
TMHMM
MEMSAT2
Transmembrane Segments
PSIPred
Secondary Structure
Functional Domain Prediction
SuperFam/SCOP
Pfam
TIGRFam
BLASTProDom
HMMSmart
Blimps/BLOCKS
FprintScan
PROSITE
Catalytic Site Atlas
ProfileScan
Functional Assignment
Genome Ontology
Enzyme Commission
Features
Manual Curation
Attributes
Web Browser
PATRIC Database

Reference Genomes Curated for the Dec. 22, 2005 Release.

Pathosystem	Complete Genomes	Genome Size	Reference Genomes	DNA curation	Protein Curation
Bacteria					
Brucella	4	3.2 Mb	1	Y	auto
Coxiella	1	1.9 Mb	1	Y	auto
Rickettsia	7	1.2 Mb	1	Y	auto
Viruses					
Calicivirus	64	7.5 kb	13	Y	N
Coronavirus	163	31 kb	16	Y	N
Hepatitis A	16	7.4 kb	1	Y	N
Hepatitis E	48	7.2 kb	5	Y	N
Lyssavirus	12	11.9 kb	1	Y	N

Pathway Analysis: plan

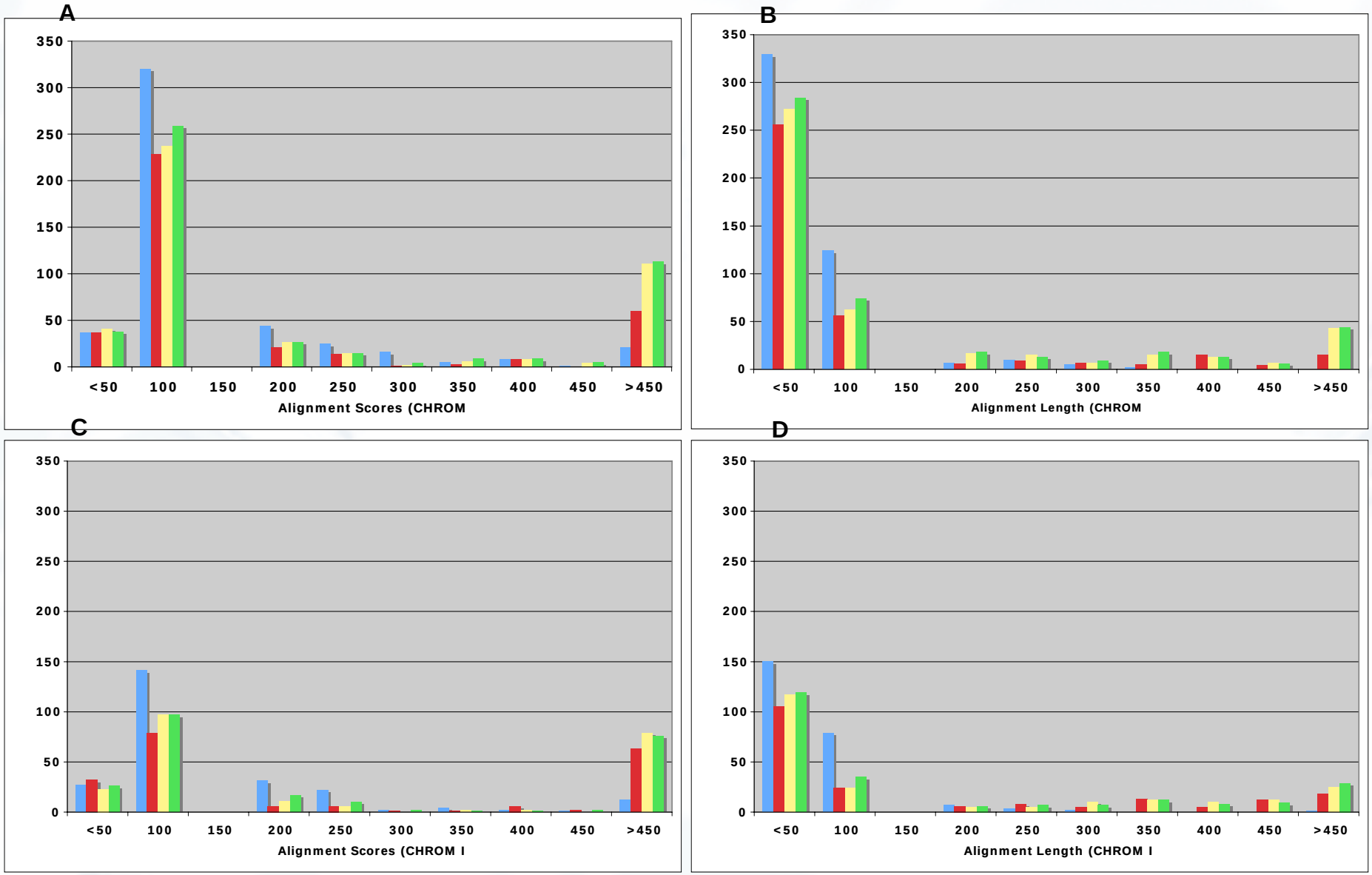
- Metabolic pathway prediction/visualization hyperlinked to protein annotation pages (Pathway Tools, P. Karp)
- Signaling pathways, protein complexes, gene regulation pathways also to be incorporated
- Host/Pathogen interaction visualizations in development
- All pathway proteins hyperlinked to proteomic experimental data

A Multi-Task Procedure for Critical Analysis of Sequence Variants and Comparative Genome Analysis

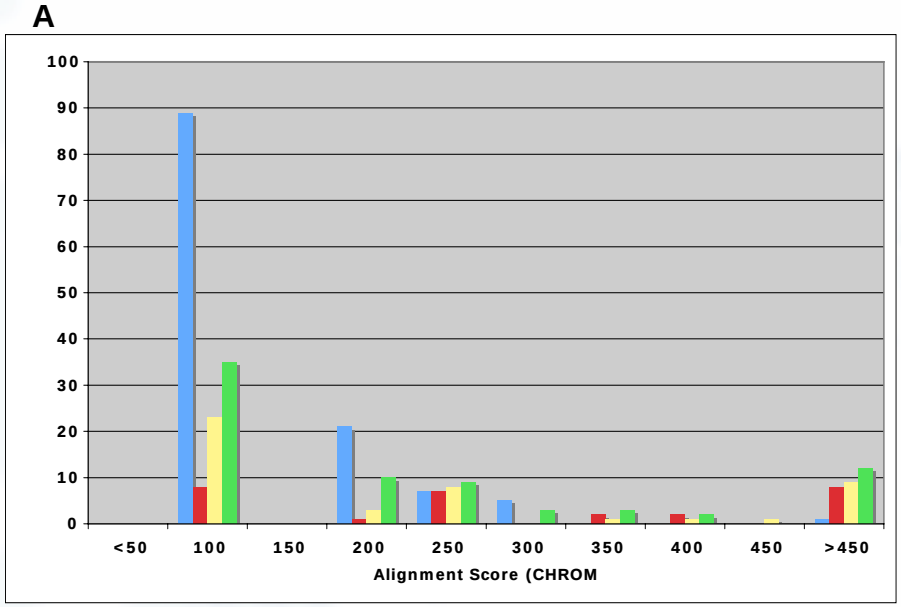
1. To identify missed genes, genes that are involved in frame-shifts, in-frame stop codons, insertions, and deletions

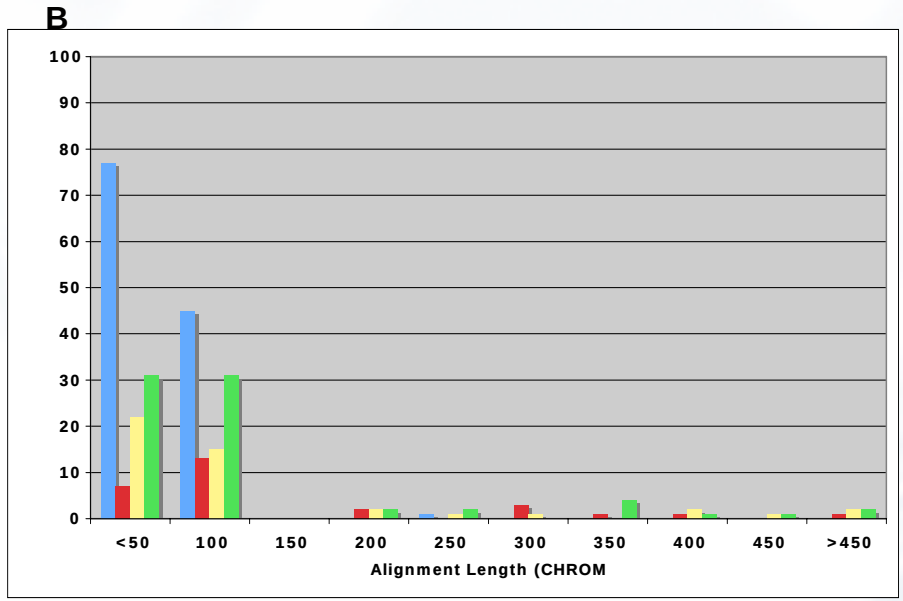
2. To comparatively analyze these gene variants among closely-related genomes to identify genome-specific genetic factors

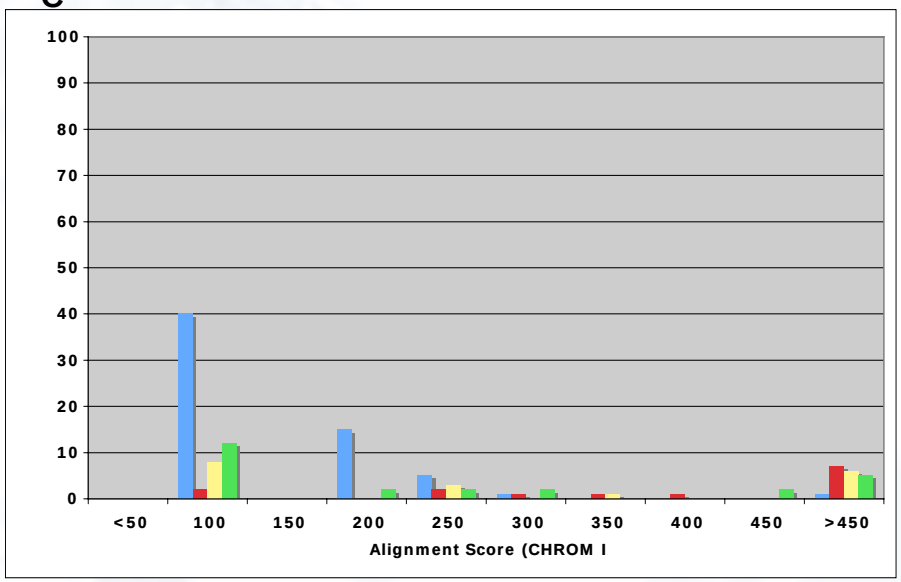
3. Build as a part of our automatic genome annotation pipeline and offer as open source software for research community.

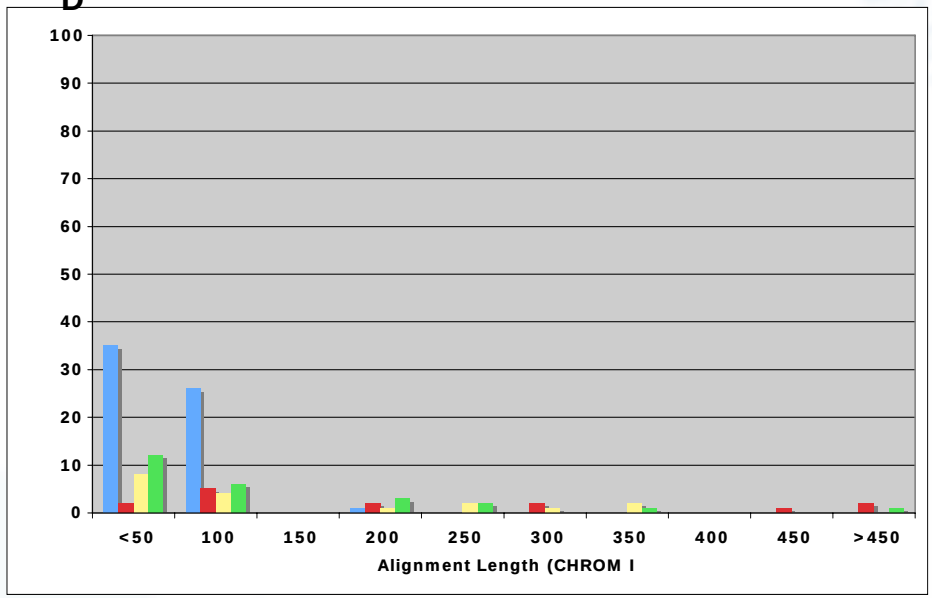


Distribution of homologs revealed in the inter-genic regions of four Brucella genomes: *Brucella melitensis* 16M (Blue), *Brucella suis* 1330 (red), *Brucella abortus* str. 9-941 (yellow) and *Brucella abortus* str. 2308



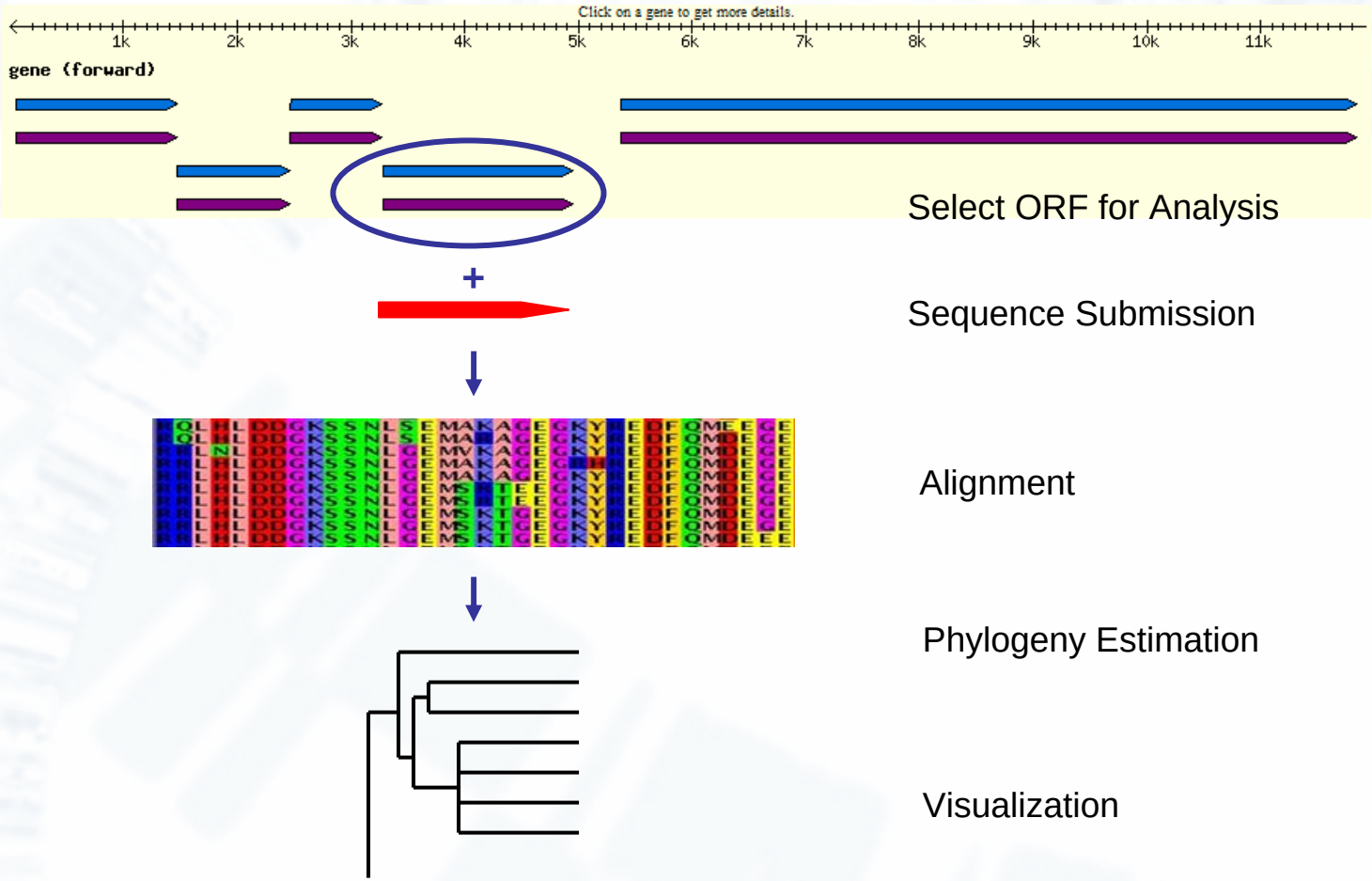






Distribution of missed genes revealed in the inter-genic regions of four Brucella genomes: *Brucella melitensis* 16M (Blue), *Brucella suis* 1330 (red), *Brucella abortus* str. 9-941 (yellow) and *Brucella abortus* str. 2308

Phylotyping Application Workflow



Select ORF for Analysis

Sequence Submission

Alignment

Phylogeny Estimation

Visualization

General plans for near future

- Steady state curation
  - Next release in June 2006
    - Nucleotide-level automated curation for all genomes
    - Automated protein-level for all reference genomes
    - Manual protein-level curation of one bacterial genome
- Various improvements to infrastructure
- LANL sequence annotation
- Special projects

PATRIC Representation in Community Meetings

Meeting	Month	Location
Rickettsia	June 2005	Spain
Nidovirus (incl. Coronaviruses)	June 2005	Colorado
Int'l Union of Microbiological Societies	July 2005	California
Int'l Virus Database Meeting	June 2005	Missouri
Brucellosis	October 2005	Mexico
Rabies	October 2005	Canada
Genome Informatics	October 2005	CSHL, NY
Biocurator	December 2005	California

Meeting Activities

- Abstracts
  - 4 Posters
  - 4 Presentations
- Sponsored Booth at IUMS
  - Cosponsored with VBRC
- Web-based questionnaires for four meetings
  - To gauge the bioinformatic needs of community
  - Nidovirus: 18 responses
  - IUMS: 9 responses
  - Rabies: 3 responses
  - Brucellosis: 0 responses

Survey Responses

Resource	IUMS Response No.	IUMS Mean Score	Nidovirus Response No.	Nidovirus Mean Score
Universal Primers	7	**2.7**	13	**1.8**
Epitope Database	7	**2.7**	14	**2.4**
Viral gene expression and assembly pathways	8	**2.8**	14	**2.4**
Phylogenetic trees	7	**3.0**	13	**2.0**
Comparative Genomics	7	**3.0**	12	**2.3**
Affect of viral proteins on host pathways	8	**3.1**	14	**1.6**
Modelling of active sites of enzymes	7	**3.3**	12	**2.4**
High-throughput data	7	**3.3**	12	**2.5**
Epidemiological data	7	**3.3**	13	**2.5**
Literature collection	7	**3.3**	14	**1.6**
3D structural data	8	**3.5**	14	**1.9**

1=Highest Priority 5=Lowest Priority

Follow Up to Survey

- Difficult to gauge needs
  - Low response to questionnaires
- Different needs of different communities
- Gap between bioinformatics resources and workflow of countermeasures development
- Approach:
  - Expedite response to one community through Special Projects; results directly applicable to one community
  - Resultant Use-Case should help build broadly applicable analysis resources

SWG meetings

- Face-to-face June 2005
- Conference call January 2006
- Next one: face-to-face June 2006